Attorney Docket No.: 17887-007600US

PATENT APPLICATION

TRAFFIC MONITORING SYSTEM WITH TRACKING BY CATEGORIES AND TERMS

Inventor: Janice Yoo
64 Van Buren Street
San Francisco, CA 94131
(U.S. Citizen)

Kian-Tat Lim
379 Everett Avenue
Palo Alto, CA 94301
(U.S. Citizen)

Stanley Ben Wong
1570 McGregor Way
San Jose, CA 95129
(U.S. Citizen)

Elliot Yasnovsky
21702 Lindy Lane
Cupertino, CA 95014
(U.S. Citizen)

Assignee: Yahoo! Inc.
3420 Central Expressway
Santa Clara, CA 95051
(a California corporation)

Entity: Large

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 415-576-0200

TRAFFIC MONITORING SYSTEM WITH TRACKING BY CATEGORIES AND TERMS

FIELD OF THE INVENTION

The present invention relates to a method and apparatus to provide statistical measurements relating to traffic served by a server or a set of servers where the traffic relates to particular topics, terms or categories.

BACKGROUND OF THE INVENTION

A server is a computing device that responds to requests from clients. A Web server is a server that is connected to the global internetwork of networks known as the "Internet" and that responds to requests received from Web clients over the Internet. As used herein, the term "Web server" may also refer to a plurality of servers organized to handle a large number of requests for a Web server, i.e., a distributed Web server system. The term "Web site" is often used to refer to a collection of Web servers organized by a business entity or other entity for their purposes. The term derives, most likely, from the language used to access one of those Web servers. A user is said to "go to a Web site" when the user directs his or her Web client to make a request of one or the site's Web servers and display the response to the user, even though the user and the Web client do not actually move physically. The user perception is that there is a location on the Web where this Web site exists, but it should be understood that the term "Web site" often refers to the Web server or servers that respond to requests from Web clients, even though "site" does not necessarily refer to the physical location of the Web servers. In fact, in many cases, the servers that serve up a Web site might be distributed physically to avoid downtime when local outages of power or network service occur.

The term "Web site" typically refers to a collection of pages maintained by a common maintainer for presentation to visitors, whether the collection is maintained on one physical server at one physical location or is distributed over many locations and/or servers. The pages (or the data/program code needed to generate the pages dynamically) need not be created by the common maintainer of the collection of pages. In places herein, such a maintainer of the collection of pages is referred to as the Web site operator. As an example, an online merchant might set up a Web server with a collection of pages created by the merchant or obtained from affiliates, suppliers or partners of the merchant and then put hyperlinks in the pages such that a visitor can browse around the "site" as

expected by the merchant. As another example, an individual dedicated to dispensing information about opera or an uncommon medical condition might set up a Web server and populate it with pages about their topic of dedication, including such things as references to pages outside their collection of pages, dynamically generated pages of comments made by visitors or e-mail sent to the operator of the Web server.

While many Web sites are targeted to single topics, some Web site operators serve many different interests and have integrated many different "properties" into a large Web site, often distributed over many servers and locations to handle traffic from a large number of visitors. For example, the Yahoo! Web site (initial URL: www.yahoo.com) brings together many properties of interest under one umbrella, including such properties as a financial property (for providing stock quotes and other financial information and data), a sports property (for providing sports scores and news), an auction property, a chat property, an instant messaging property and many others. Such sites, where visitors come for possibly unrelated properties, are often referred to as "portal sites".

While the typical Web site includes one or more servers that receive requests and provides responses according to the HyperText Transport Protocol (HTTP), the description herein should not be understood as being limited to a particular protocol or a particular network. For example, the Web site might be connected to the Web clients via an intranet, wireless access protocol (WAP) network, local area network (LAN), wide area network (WAN), virtual private network (VPN) or other network arrangement. In other words, a Web site for which traffic is being monitored can be monitored independent of the protocols or network used. "Web" typically refers to "World Wide Web" (or just "the WWW"), a name given to the collection of hyperlinked documents accessible over the Internet using HTTP. As used herein, "Web" might refer to the World Wide Web, a subset of the World Wide Web, a local collection of hyperlinked pages, or the like. More generally, a Web server is a server responsive to requests received from a Web client.

Typically, requests and responses are considered "pages". For example, with the HTTP protocol, a Web client requests a page from a Web server and the Web server responds to the request by sending a page. In the HTTP protocol, a Uniform Resource Locator ("URL") identifies a page and that URL is presented to the Web server as part of a request for a page. The pages are often HyperText Markup Language (HTML) pages or the like. The HTML pages can be static pages, dynamic pages or a

combination. Static pages are pages that are stored on the server, or in storage accessible by the server, prior to the request and are sent from storage to the client in response to a request for that page. Dynamic pages are pages that are generated, in whole or in part, upon receipt of a request. For example, where the page is a view of data from a database, a server might generate the page dynamically using rules or templates and data from the database where the particular data used depends on the particular request made.

The term "page hit" refers to an event wherein a server receives a request for a page and then serves up the page. In even a moderate sized Web site, the servers might handle millions of page hits per day. A common measure of traffic at a Web site is in the number of page hits (often referred to as "page views", especially in an advertising context) for particular pages or sets of pages. Page hit counts are a rough measure of the traffic of a Web site. More refined measures include unique visitor counts, where only one page hit is counted for each unique client per some period. Such measures work well when the traffic of interest relates to particular pages, but are generally not informative when traffic by topic is desired and multiple pages may relate to one topic and one page may relate to multiple topics.

For example, where a stock information Web server just serves up a page for each stock and only one page relates to that stock, it would be a simple matter to determine levels of user interest in particular stocks by just examining the server logs of the Web server to determine which stock pages are being served the most. Unfortunately, most real-world Web services are not so well defined. One more complex Web site includes servers that serve news, sports and financial content along with content on many different subjects and pages that relate to a common topic might be served from more than one of those content components. With the requests spread over different content components, the level of user interest would not be accurately reflected in just a measurement of interest in one content component. For example, interest in a particular athletic shoe company might be expressed by traffic to pages containing news stories relating to the company, traffic to sports pages referring to the company, traffic relating to financial content about the company, searches for the company's products, purchase transactions for the company's products, etc. Also, some requests might be falsely associated with interest in the company if, for example, users use a search term that has more than one meaning, one of which relates to the name of the company.

Such a Web site might also include search capability, wherein a user submits a search request using their Web client and a Web server responds with a page

that contains search results. It is a simple matter for a search engine (a Web site set up to respond to search requests) to log all of the search requests. Typically, a search request is in the form of a search phrase containing one or more search terms. Search requests can be counted by search term, e.g., count the number of times "Ford" or "sports" was used as a search word in a search phrase, but such counts have limited utility where one search term might relate to multiple topics and multiple search terms might relate to one topic.

Where page hits, search requests, or other "events" such as purchases, are logged or loggable, some operators of Web sites track statistics other than just page hits or search requests. One well-known statistic that is often seen in Web systems, and elsewhere, is a "top-*n*" list, such as a "Top Ten" list. Such a list presents the *n* highest requested items. For example, a newspaper might list the 40 best selling books for a given month, ranked by industry-wide sales. The list might indicate, for each book on the list, the book's ranking for the prior measurement period. As another example, a Web site operator might include a page served by their Web server(s) that lists the top sellers for that operator.

As yet another example, a Web site operator might include a page served by their Web site that shows the top sellers for various categories. For example, if the Web site operator is a toy retailer, the operator might create pages to be served by their system wherein the pages list the top selling toys for infants, the top selling toys for infants, the top selling toys for teens, etc. In a variation on the basic count of items sold, some Web site operators might include statistics showing how various items are moving up or down in sales. For example, a list could be presented showing the top 40 sellers for the month along with their sales rank for the prior month, or a list ranking items in order of increase in sales or sales rank.

As with the Web server that serves up specific pages for specific topics, such as one page per stock on a stock information Web site, sales statistics such as those described above are easy to generate. An electronic commerce server can simply log each purchase and then a program can scan the log for a period of time to determine sales levels for each item. The sales can also be categorized easily where the items are already categorized. For example, a book selling Web site can log all sales of books, where each book is already categorized (e.g., "fiction", "reference", "technical", "self-help", "other nonfiction", etc.) and then aggregate the sales for category to identify sales by category or top sellers within a category. However, the 'top-*n*" or best seller lists are limited in that the categorization of the items must be done manually or along lines that are set out ahead

of time and worked into the data. Thus, such a system cannot be easily adapted to events that are not already well-categorized, it does not combine information across multiple events and types of events, nor is the information normalizable so that detailed and relative statistics can be derived.

Some traffic analysis modules have been used to analyze traffic over a Web site, but their functionality is limited. One such module performs basic statistical analysis of Web server logs to determine Web site usage. They are typically not designed to compute interest in particular topics, although the statistics they offer indirectly reflect that interest. One problem with such modules is that they either rely on manual associations of events to topics or they do not associate events with topics, so the former approach is not scalable and the latter approach does not group events in a meaningful manner.

Heretofore, however, none of the statistics systems described above allows for the more sophisticated, and thus informative, measurements often needed to make overall strategy decisions with regard to trends, advertising purchases, popular culture review, product marketing and other decisions that need to be made in light of traffic statistics where the traffic relates to complex events and requests.

SUMMARY OF THE INVENTION

Using the present invention, a traffic monitor generates statistics about traffic of one or more servers and is capable of associating monitored events with topics or terms and aggregating the statistics about the monitored events into categories. One use of such statistics is to determine trends and changes in interest, in effect detecting "buzz" due to increased interest, where such interest is associated with a topic, term or category.

In an alternate embodiment, instead of monitoring traffic resulting from requests from any set of users to a specified set of Web servers or Web sites (operated by one or more entity), the traffic between a defined set of users to any set of Web servers could be monitored instead.

Monitored events might include page hits, search requests, purchases and/or other actions. In one embodiment of a traffic monitor, events are associated with topics or terms and are grouped by category. For example, where a user provides a search server with search terms and then selects a page from search results, the resulting page hit might be associated with one or more of the search terms used. Where a user arrives at a

particular page after navigating a subject directory, the page hit might be associated with the subject of the navigation. By comparing changes or trends in the traffic associated with a search term or a category, the "buzz" associated with a topic, term or category can be assessed.

In a process of evaluating traffic, the raw values can be normalized to reduce the effects unrelated to the buzz around a topic, term or category. For example, while raw values for traffic are likely to grow from midnight to midday in a given geographical area as users awake and begin accessing the server system, the traffic measurement can be normalized to remove time of day variations. Other variations, such as overall traffic variations, seasonal variations, weekly variations and general topic variations (when examining buzz for more specific topics), can also be normalized out. Ratios and difference measurements might also be performed in comparing two or more topics, terms or categories to determine relative buzz.

Once "buzz" (a statistical measure of interest) is determined for a set of topics, terms or categories, that information can be used in many ways. For example, users might be interested in seeing what are the current popular terms or categories, so that they can follow trends and be informed on those popular topics. Advertisers might also be interested in buzz, as they might want to dynamically switch their advertising to follow topics having increasing buzz.

One advantage of a traffic monitor having aspects of the present invention is that the traffic monitor will group events such that a user of the statistical data can get statistics that cover events that relate to a topic without including counts for events that are not really on the same topic. Yet another advantage is that counts can be normalized for a topic or term against other topics or terms in a category.

A further understanding of the nature and the advantages of the inventions disclosed herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a Web site system including a statistical analyzer within which a traffic monitor according to one embodiment of the present invention might be used.

Fig. 2 is a graph of a category hierarchy, showing categories and subcategories, as well as terms associated with categories and subcategories.

Fig. 3 is a schematic of a data structure used to represent counts by category and topic/term; Figs. 3(a) and 3(b) show data structures.

Fig. 4 is a schematic of a data structure for storing multiple sets of traffic data, one set per period.

Fig. 5 is a schematic diagram of a canonicalization system.

Fig. 6 is a flowchart of a process for categorizing search words.

Fig. 7 is a block diagram of server system including a traffic monitor according to one embodiment of the present invention.

Fig. 8 is a flowchart of one process for generating buzz/trend reports.

Fig. 9 is an illustration of a buzz report.

Fig. 10 is an illustration of a list of vertical market topics for which buzz can be presented in the exemplary report of Fig. 9.

Fig. 11 is an illustration of a report where the buzz for terms is plotted over time and relative to other terms in a category.

Fig. 12(a) and 12(b) together illustrate a report showing buzz values for subcategories in a category.

Fig. 13 illustrates a campaign monitoring page.

Fig. 14 illustrates a campaign monitoring report.

Fig. 15 illustrates intersection analysis.

Fig. 16 illustrates associated interests analysis.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

The following description is organized approximately according to the following outline:

1. Overview
2. Collecting Traffic and Binning by Subject
    - 2.a. Categorization
    - 2.b. Canonicalization
3. Examples of Sources of Data for Traffic Monitor and Uses for Collected Data
4. Uses of the Statistical Analysis
    - 4.a. Buzz/Trend Reports
    - 4.b. Selling Advertising Space Based on Categorizations and/or Buzz
    - 4.c. Campaign Monitoring
    - 4.d. Intersection Analysis
    - 4.e. Associated Interests Analysis
5. Variations on the Basic System

1. Overview

In this description, the term "buzz" refers to a measurement of the traffic that relates to a particular topic, term or category. As used herein, "subject" generically refers to one or more of a topic, a term, or a category. Thus, the topic "U.S. presidential politics", the search term "Ford" and the category "music", are all subjects for which "buzz" can be measured.

Traffic refers to a count, or approximate count, of the events that occurred for a given subject. Traffic can either be measured for a defined set of servers accessed by a possibly unconstrained set of clients/users ("selected servers/all clients"), for a defined set of clients/users accessing a possibly unconstrained set of servers ("all servers/selected clients"), or for a defined set of clients accessing a defined set of servers ("selected servers/selected clients"). For example, the selected servers might be the servers that serve content for one or more defined Web sites, the servers that are monitored by an advertising network or ratings network, the servers monitored by a university network monitoring system, etc.

Traffic might be a raw count of the number of events, unnormalized or otherwise, but traffic might also be measured not with one count per event, but one count per unique user (i.e., even if a particular user makes multiple requests, only one request is counted) or one count per unique user per time period might also be the measure of counting traffic. Traffic can be unnormalized, such as integer counts for the number of events, or can be normalized. One purpose for normalization is to place the number in a suitable value range for presentation or other processing. Another purpose for normalization is to normalize out variability in the counts that is likely to be variability independent of levels of user interest.

In general, monitoring traffic for any users or any servers ("all servers/all clients") is only practical in a centrally managed system and cannot currently be effected for Internet clients and servers in general, however if logs of such activity were available, the traffic monitors and statistical analyzers described herein might be used to measure traffic and buzz in a more general setting. The examples herein largely refer to the "selected servers/all users" variation, but one of ordinary skill in the art would understand how to apply this disclosure's teachings of that variation to the other variations.

Events can be page views, search requests, purchases, requests for media such as streaming audio or video, message board actions, chat room actions, club actions, instant messaging actions, online gaming actions, or any other action that is detectable by

a server of a Web site. The expected use of the traffic monitor is to monitor large numbers of events, often measuring in the millions, to discern trends and buzz. To enhance the usefulness of the results, events should be logically grouped so that the groupings will by and large have statistical significance and topical relevance. The process of grouping events is referred to herein as "binning".

Whatever the extent of the traffic monitoring (e.g., selected servers/all users), the results can be sliced up by demographic information. For example, the traffic monitor can provide the overall counts for the category "music", but the traffic monitor can also divide up the overall counts by different demographic categories, using user-provided demographic data or demographic data provided in another way. For example, the traffic monitor can provide buzz values for the demographic of 18-45 males with U.S. addresses. An example of demographic information other than user-provided information is the user's client's IP (Internet Protocol) address. Examples of user-provided information include age, gender, residence location, and user preferences, such as browser type, client type, network type, etc. In addition to slicing up the data to show traffic for a particular demographic, the demographic data can be used to show how a particular count for a topic is divided up among the demographic categories.

2. Collecting Traffic and Binning by Subject

Fig. 1 is a block diagram of a traffic monitor 100 including a canonicalizer 102, a categorizer 104, a count generator 106 and a canonicalization database 108. Canonicalizer 102 is coupled to receive search log records and page hit records to determine, for a given search request or page hit, what the relevant topic is. Canonicalizer 102 might refer to canonicalization database 108 to resolve canonical terms.

In alternate embodiments, different sets of one or more server logs are used to identify the bin or bins for which counts are incremented for an event logged in the server logs. For example, the system shown in Fig. 1 might include an additional log of purchase records or streaming media downloads. Where the events to be binned are purchase events, each event can be evenly weighted or each event can be weighted according to a purchase amount.

As an example of a specific traffic collection operation, suppose that thousands of users connect to a search server and perform a search using the phrase "local weather". The search server might respond to that phrase by presenting the user with a results page including links to pages relating to weather and specifically local weather

(where locality might be inferred from user preferences or other methods). The search server logs the search itself and the "clicked-through" pages from the results page. A page is a "clicked-through" page when a user notes a reference to that page on the results page and selects that reference from the results page. In a standard HTTP system, the effect of those actions is that the user's browser (or other HTTP client) requests the referenced page from the server indicated in the reference (which might or might not be a portal server) and the referenced server responds to the request with the referenced page.

If the search server serves pages from a potentially large number of pages, tracking hits for each page might result in statistics that are too granular to be useful. Because of this, it is often useful to aggregate hits by subject. For example, if there are fifty requests for a local weather page in a day for fifty different localities, it might be more informative to state that there were fifty requests for weather information than to state that there was one request for weather in a given locality. Because of this, in the preferred embodiments, traffic monitor 222 aggregates counts into bins, where each bin is for a particular topic or term.

A given event can be binned with other events that relate to the same topic or term to achieve statistical significance and topical relevance to the counts for the topic or term. In other words, the bins contain enough counts to be statistically significant and events that really relate to the same topic are binned together, even though the events may appear to be quite different. For example, page hits for a page known to relate to the U.S. presidential elections can be binned with page hits for other pages known to relate to the U.S. presidential elections. Where the page hits are the result of a given search term, the page hits are binned with other results for the search term. In this manner, counts are accumulated for a bin associated with that topic or term. A given topic or term is associated with one or more categories. Of course, a traffic monitor could be designed wherein a topic or term might be associated with none of the categories, but it is usually best to consider that any given topic or term falls into at least one category, even if the category is a catch-all "overall events" category or the root of a category hierarchy.

In some implementations, categories are organized hierarchically, with a first level of categories, subcategories within categories and possibly subcategories within subcategories. In this hierarchical arrangement, an example of which is shown in Fig. 2, topics/terms are associated with categories and/or subcategories. Unless otherwise indicated where "category" is used herein, it should be interpreted to refer to a category or a subcategory.

In some cases, one topic/term is present in more than one category, as with (see Fig. 2) the term "New Orleans", which is found in the categories "Music", "Blues" (itself a subcategory of "Music"), and "Travel". Typically, where one term is present in two or more categories, the term has two meanings. If the meaning can be discerned from the context, then only the count for the actual meaning of the term should be incremented. In the following section, a categorizer for identifying the particular bin or bins in which to count an event is described. For example, if the context of an event was travel to New Orleans, the count for the term "New Orleans" under the category "Travel" would be incremented, but the counts for the other "New Orleans" terms would not be.

Fig. 3 illustrates one possible arrangement of data structures for storing the counts for bins. As shown in Fig. 3(a), a category record 150 contains data elements relating to a label for the category and counts for each of a plurality of demographics. Where demographics are not used, the category record would just store a single count. Since counts are for topics and terms, the category record need not contain the count(s) for the category. Instead, the count(s) for a category could be determined by summing the counts for all the terms that are in that category. However, in system with large numbers of events, storing the counts for categories may result in a much faster system than if the category counts had to be calculated each time they were used.

Also shown in Fig. 3(a) is a subcategory record 152. Subcategory record 152 is similar to category record 150 except that subcategory record 152 includes a pointer to the category for that subcategory.

Fig. 3(b) illustrates a bin record 154 associated with a topic or term. Bin record 154 includes a label for the topic or term and includes count data for one or more categories (or subcategories). For each represented category, bin record 154 holds count data for that topic/term in that category as well as a pointer to the category.

Fig. 4 illustrates a data structure 170 that might be used to store multiple sets of traffic data, one set per period. In this example, the period is daily, so data structure 170 stores a collection of category/subcategory records and bin records for each of a plurality of dates.

2.a. Categorization

Categorizer 104 determines the bin or bins that have their count incremented for a particular event. For example, where the event is a search request using the search phrase "formula one" and the search results page lists pages related to algebra and auto racing, the search might be categorized under mathematics or sports.

However, categorizer 104 correlates searches with search results selected, so that when the logs show that the user selected from the search results a page relating to auto racing, categorizer 104 allocates that event to the "auto racing" category and the "formula one" term in that category. Where terms remain ambiguous even after selection of a page (or if the user does not select a page from a search results page), categorizer 104 might output fractional counts for more than one category with suitable weights summing to one.

In some cases, the category associated with a page hit or a search are readily determinable by the state of a visitor's server session. For example, if the user is navigating a search directory by category/subcategory using a search term and then selects an entry under a subcategory, then the count for that event is readily allocable to the bin for the search term under the category and/or subcategory previously assigned to that entry. For example, if a user navigates the Yahoo! search directory path "Top: Sports: Regional Sports: San Jose" using the search term "scores" and selects a page from the result, then the categories and subcategories that get the count are readily ascertainable.

However, with direct searches with words having multiple meanings, the category might not be so apparent. For example, if the user started a search within the Yahoo! search path "Top:" and requested a search on "Ford" and "Michigan", the category is unclear because the visitor might be interested in the Gerald R. Ford Library in Ann Arbor, Michigan, or the visitor might be interested in the Ford Motor Company, which has offices in Michigan. One method of resolving the ambiguity is to examine the resulting clickstream. For example, a Yahoo! search directory search using the search phrase "Ford Michigan" might return several matches, including those shown in Table 1.

Table 1

```
Regional > U.S. States > Michigan > Cities > Ann Arbor > Education >
College and University > Public > University of Michigan > Libraries and
Museums
     Gerald R. Ford Library

Regional > U.S. States > Michigan > Metropolitan Areas > Detroit Metro >
Business and Shopping > Shopping and Services > Automotive >
Dealers > Makes
     Ford
```

When a user is presented with the entries shown in Table 1 and selects the first clickable link (Gerald R. Ford Library), the categorizer would assign the count for the event to the "Libraries and Museums" subcategory (and to each higher level subcategory if such

tracking is performed). However, if the user selects the second clickable link, the categorizer assigns the second category/subcategory path shown in Table 1.

Where the categories tracked by the statistics monitor overlap the category structure of the search directory, the task of assigning counts is complete. However, where the structure of the statistics monitor does not overlap the structure of the search directory, some additional steps might be performed. For example, if the statistics monitor had categories for each U.S. state and categories for each U.S. President, then the count for the search term "Ford Michigan" followed by a click on the first clickable link in Table 1 might result in the statistics monitor assigning half a count to the category for Michigan and half a count to the category for former U.S. President Gerald R. Ford.

In a more precise implementation of such a system, the counts might not be even. Continuing with the example of Table 1, more than half a count might be assigned to the more likely category of interest and the remainder to the other category. Thus, one might expect that a click on a link to the Gerald R. Ford Library is more likely to reflect an interest in the library as opposed to an interest in Michigan, where the library happens to be located.

The search engine for the search directory returns a list of matches with one or more clickable link per match. Generally those links can be categorized into one of three categories: 1) internal pages, 2) external pages categorized internally and 3) external pages not categorized internally. A Type 1 link is easily categorized by assigning a category to the page pointed to by the link. A Type 2 link does not have an explicitly assigned category, but can be categorized because the referenced page is referenced elsewhere on the portal site by a Type 1 link. The categorization for Type 1 links is easier than categorizing all possible search terms, and may have already been done if the search directory is organized by subject, as with the Yahoo! search directory.

Fig. 5 illustrates one process to categorize search words for Type 1 and Type 2 links based on the link selected. Type 3 links can be binned as well, if some category indication is present or a categorization engine that handles such links is used to identify their categories. As shown, a categorizer would extract the search word events (user ID, timestamp, search words) from search logs. The user ID can be implemented as a unique cookie stored in the user's Web browser that is sent to the search engine and Web page server with each request and is stored in the logs.

The categorizer also extracts from page view logs the user ID, timestamp, page ID, etc. for each page view. After sorting both of the extractions by time, the

categorizer can interleave the extractions and determine which page is viewed after a user views results of his or her search. From that determination, the categorizer can look up the category of the viewed page and that category can be attributed to the search. Where the search is being tracked for buzz evaluation or other counting evaluation, the category count is incremented. Where a category cannot be determined, the event can be ignored for monitoring purposes.

In previously developed categorizers, the search terms are used to identify the category that gets credit for the hit, but using the above method, the category is identified from the page that is visited after the search, eliminating the need for complex semantic analysis to resolve ambiguities or manual categorization of search words, which is not scalable to a large system.

As an alternative to the method described in Fig. 5, the links on the search result page can be rewritten to include "redirects" (i.e., intermediate commands executed upon a click) that log the page ID and search phrase, so that only one log is needed. With one log, the sorters and interleaver are not needed.

Either way, the categorizer finds the meaning of a search term that the user ascribes to the term, in an inherently scalable way.

2.c. Canonicalization

When dealing with search words, it often makes sense to combine information about similar terms that are intended to produce the same results. For example, a term may be misspelled, or it may have words in a different order than another, or it may contain non-essential words such as "the". The process of reducing such terms to a common, standard form is known as canonicalization. Many processes are known for performing canonicalization, ranging from less aggressive processes such as removing certain punctuation characters or so-called "stop words" such as "of" and "the", to more aggressive processes such as adding, changing or deleting letters within words.

The canonicalization process might be performed by canonicalizer 102 that is part of traffic monitor 100 (see Fig. 1). As an example, canonicalizer 102 might canonize the search phrase "Denver whether" to "weather" by inferring that a spelling error occurred. As with categorizer 104, canonicalizer 102 uses user behavior to improve the canonicalization process. Using user behavior is inherently scalable because there are generally proportionately more users to give human input as the system grows larger to handle more traffic.

Using user behavior (a large increase in number of searches) also allows more aggressive canonicalization. For words whose search usage has increased rapidly, more aggressive canonicalization techniques can be used. In addition, when combining information (such as number of searches) about such aggressively canonicalized terms, the system does not just add the values, but transfers the portion of the value that exceeds a prior baseline value to the canonicalized term, leaving the remainder attached to the raw, uncanonicalized term. For example, if "Concord" (Massachusetts) has a current value of 420 and is to be combined with "Concorde" (the airplane) with a current value of 825, and "Concord" had a prior baseline value of 130, we transfer a value of 290 (420 - 130) to the canonicalized term, ending with "Concord" at 130 and "Concorde" at 1115.

The baseline value can be defined as the average of the value for a previous period. In one embodiment, the baseline value is retained. If the value for the term being combined declines to its previous baseline, the terms are no longer merged. Combining only values over baseline more accurately reflects reality for terms with multiple meanings.

Fig. 6 illustrates a typical implementation of a canonicalization process. The aggressive canonicalization step might include adding, changing or removing letters from search terms. If the value of the term being merged is within some margin, such as 20%, of its baseline, the term is no longer merged. Terms (or fractions of the values of terms) should be merged when they are likely to be about the same topic. In the case of rapidly changing terms, it is unlikely that two similar-appearing but conceptually different terms will both have rapid rises at the same time. Thus, it is possible and desirable to merge similar-appearing terms that both have rapid rises, since they most probably relate to the same concept or topic.

For example, the term "U.S. Open" might exhibit rising interest. If the term "U.S. Open Golf" is also exhibiting rising interest, but the term "U.S. Open Tennis" is not, the canonicalizer assumes that term "U.S. Open" and "U.S. Open Golf" refer to the same subject and can be combined but "U.S. Open" and "U.S. Open Tennis" should not be combined. Once the interest levels in "U.S. Open Golf" or "U.S. Open" fall back to around their baseline, the canonicalizer would separate these terms out again, to have them binned separately. This would provide a desirable system response, at least for the above example, because depending on the timing of the U.S. Open sporting events, "U.S. Open" might relate to "U.S. Open Golf", then fall back near its baseline and then rise

along with "U.S. Open Tennis", at which point "U.S. Open" would be associated with the "U.S. Open Tennis" category.

Thus, the canonicalizer would respond to canonicalizations that change over time, as is often the case in the real world of user interests. When combined with other elements of a traffic monitor, the buzz values for terms that reflect actual user interests are readily available for use by the canonicalizer to determine which topics/terms to merge and when.

3. Examples of Sources of Data for Traffic Monitor and Uses for Collected Data

Fig. 7 is a block diagram of server system 210 including traffic monitor 100 according to one embodiment of the present invention. In server system 210, users connect to servers 214 by connecting user computers 212 to servers 214 via a network 216. In a specific implementation, user computers 212 are Internet-connectable computers (desktop computers, laptop computers, palm-sized computers, wearable computers, set-top boxes, embedded TCP/IP clients, and the like), servers 214 are Internet-connected servers responsive to requests at a URL designated by the portal operator and network 216 is the "Internet". The typical computer 212 includes a browser or other HTTP client that is used to provide a user with HTTP access to the Internet and the Web.

The particular details of how a particular user computer 212 connects to a particular server 214 and how the particular server 214 is selected are not shown here, as there exist many such arrangements and the present invention is not limited to any particular client-server arrangement. In the figures, distinct instances of like objects are distinguished with parenthetical indices. For example, user computer 212 might refer to 212(1) or 212(n). As used herein, "n" refers to an indeterminate integer where the actual value of the integer is not relevant and may depend on details not relevant here. It is used in various contexts and the value of "n" may be different in each context, unless otherwise indicated. For example, the user computers in Fig. 1 are shown ranging from 212(1) to 212(n) and the servers are shown ranging from 214(1) to 214(n). Thus, one can infer that there are an indeterminate number of user computers and servers and the actual number is not relevant for the purposes of this description, but one should not infer that the number of user computers and servers must be the same.

Fig. 7 shows, in addition to user computers 212 and servers 214, several other components, such as storage for server logs 220, traffic monitor 100 with inputs for reading server logs 220 and outputs for count data to be added to a statistics database 224.

Also shown is a Web server 230 coupled to network 216 and a database server 232, that is in turn coupled to statistics database 224.

In a typical operation, a user connects a user computer 212 to a server 214 and requests one or more pages, with each page being identified by a URL. Because of the user perception of this process, it is often described as a visitor visiting going to a particular page on a Web site as defined by a URL, to analogize to physical movements. However, the visitor does not actually move anywhere and there might not be a physical "site" that can be pointed to as the place that is visited. Nonetheless, such analogies have become quite common and are used herein. Thus, it should be understood that the act of a user or "visitor" going to a page on a site is normally an act of the user or visitor directing its computing device to make a request through a network that handles such requests, wherein the request is for a page specified by the URL of the request and maintained on a server specified in the URL or the request, along with the act of receiving a response from the server and possibly displaying it or processing it.

In current use, even the term "page" is somewhat of an analogy to the beginnings of the World Wide Web, when the requests were for page files stored in directories on the server specified in the URL. However, in current use, "page" refers to what is returned by the server and thus a page might be data that is not even in existence at the time of the request (e.g., dynamic Web pages).

One possible order of events will now be described with reference to Fig. 7. The events described below correspond to circled numbers in the figure which are parenthetically referenced in the text below. One of ordinary skill in the art will recognize, after reading this disclosure and review of the figures, that other orders of events are contemplated by this disclosure and many equivalents can be inferred from the figures and text.

The events illustrated by the circled numbers begin with a process of logging page hits (1) occurring on servers 214. Many of the details of the logging process are described in further detail above. Once the server logs are created, traffic monitor 222 can read the logs to identify counts of hits by subject (2) and store those counts in statistics database 224 (3). The next event is where a user issues a statistical query relating to buzz (4). As shown, the user issues the query using user computer 22(a), but it should be understood that any computer or computing device with sufficient rights and capabilities, including user computers 212(1) through 212(n), could be used for buzz queries. From whatever source, Web server 230 receives a request for buzz statistics and

translates the request into a database query, which is presented to database server 232 (5). In response, database server 232 reads data from statistics database 224 (6) and returns a database result to Web server 230 (7). Web server 230 then formats the database result into a Web page and delivers that page to the requesting user computer (8). An example of such a delivered page is the page shown in Figure 8. That example page is responsive to a request for top buzz values for overall events and events specific to the categories of movies, music and sports.

Note that, depending on the device making the request, what is returned by Web server 230 might not be in the form of an HTML page, but would typically be in a form usable by the requesting device. For the purposes of providing at least one specific detailed example, this description assumes that user computers 212(1)-(n) are HTTP clients and request pages interactively from servers 214, that user computer 212(1) is also an HTTP client and interacts with Web server 230 in a conventional manner. While it should be understood that where many querying devices are in use, Web server 230, and possibly database server 232, might be replaced with arrays of servers to handle the load of statistical queries.

In one embodiment of a traffic monitor that is described herein, the monitor operates off of usage logs generated by a Web site's servers. Notwithstanding that description, it should be understood that the monitor might operate off of other indications of traffic, such as real-time page hits, click streams, purchase records or database records. Furthermore, while the traffic monitor is shown as a unified system, a distributed traffic monitor might be used where such distribution aids in making the traffic monitor scalable and less computationally complex, all without necessarily departing from the scope of the invention. It should also be understood that the present invention is not limited to a particular Web site or collection of Web sites, although many of the examples show examples from a specific Web site, namely the Yahoo! Web site.

4. Uses of the Statistical Analysis

As described above, a "buzz" value represents the level of interest of a subject, such as a movie, a person, product, place, or event, cultural phenomena, etc, and the change in buzz value might be indicative of a trend. The buzz value can be calculated as the number of unique users searching for that subject anywhere on a portal site or set of portal sites or viewing a page of content relevant to that subject anywhere on the portal site or set of portal sites. As described herein, buzz might also be calculated without regard to whether each event that is counted is originated by a unique user.

The buzz values can be used to identify cultural trends, track interest in specific brands, measure the effectiveness of marketing campaigns, etc. For buzz events that are purchase events, the count by which a bin is incremented might be a function of purchase amount, so that purchases of larger amounts have more of an effect on a product's buzz than purchases of smaller amounts.

In one variation, the buzz value associated with a particular term or category is the number of users searching on that term, or viewing a page related to that term, divided by a sum of users searching, where the sum can be the sum of users searching over all subcategories in a category, sum of users searching over all terms in a category, or sum of all users searching anywhere on the site. The latter normalization is useful to factor out time-based increases in traffic, such as weekday-weekend patterns, seasonal patterns and the like. A normalization factor might be applied to all terms being compared so that the buzz values are easily represented. For example, if there are four terms in a category, 100 total unique user hits on those four terms (25, 30, 40 and 5, respectively) out of one million total unique users, a normalization factor of 100,000 might be applied so that the buzz values are 2.5, 3, 4 and 0.5, instead of 0.000025, 0.00003, 0.00004 and 0.000005. Normalization can also be used when determining the buzz surrounding one company or product against an index of other companies or products within a particular market segment or product category.

In some cases, the buzz values for a subject might be a leading indicator for electronic commerce transactions relevant to that subject. For example, the buzz for the term "widget" might rise and be followed by increased on-line purchases of widgets. Such information is useful to advertisers interested in having their brand of widgets be selected, as well as fulfillment managers eager to have in stock the latest trendy items.

Buzz values can be presented from overall data or it can be isolated to specified demographic groups. Thus, with enough traffic, the traffic monitor can track the top buzz among women aged 33-45, the top buzz for "newbies" (people who are new to the online world), buzz by country, by regions of countries. In addition to just a buzz number, the system might also provide a commerce index to show how different vertical markets or products are growing/shrinking over time.

While advertisers and other businesses might find the buzz values to be useful and key marketing feedback data and thus be willing to pay for the data, other buzz values might be made available to consumers or the public in general for free or a nominal cost. For example, the Web site operator might opt to provide general access to

the buzz relating to current movies and rock stars while providing more restrictive access to data relating to a particular marketing campaign being tracked by the operator for the company that launched the campaign.

4.a. Buzz/Trend Reports

Fig. 8 is a flowchart of one process for generating buzz/trend reports.

One example of a buzz report is shown in Fig. 9. That report has a section for buzz values (normalized from the counts) for overall terms as well as sections for the categories of movies, music and sports. For each section of the report, the report shows the top few topics/terms that generated the most counts, in order of number of counts, along with an indication of relative change in buzz values. When implemented as a hyperlinked page, the report also includes links to a list of categories (the link is denoted by "A" in Fig. 9 and the "linked-to" page is illustrated in Fig. 10), icons to change the sort order, as well as related links related to the particular topic/term (e.g., news, categories and sites relating to the topic/term).

As shown in Fig. 10, while the counts might be separated in a bin record by category (see Fig. 3(b)), the counts can also be aggregated over all categories.

Another buzz report is shown in Fig. 11, where the buzz for terms is plotted over time and relative to other terms in a category.

Fig. 12 illustrates yet another buzz report, showing buzz values for subcategories in a category. As shown the subcategories are for the category "Music" and are sorted by percentage change in buzz value. Fig. 12(a) and 12(b) together form the report. In the portion of the report shown in Fig. 12(b), the buzz for terms over the category "Music" are there shown. A link to a customization page is provided ("Preferences") as well as a link to a user-specific buzz index ("My Buzz Index").

In general, there are many ways to present the data generated by the traffic monitor. Buzz values can be "sliced" by demographic to illuminate demographic information about the users searching for a particular search term. Buzz values might be sliced by method of access, such as wireless or broadband access. Buzz values can be presented in various sort orders such as "buzz score" or by the "% change in buzz" for the time period specified. Users of the buzz reports can easily determine, for some demographic or overall, what topics or search terms get more attention and where the spikes in attention occur over time.

In one application, a buzz report generator generates buzz reports on the fly based one requests from users of the buzz report generator. Thus, such users can

request and receive customized views of buzz by segments. A buzz report generated by the buzz report generator can be presented for any type of user segments that can be defined by user characteristics such as demographics, lifestyles, interests and/or geographic location.

Demographics of users can also be used as added data rather than just as a way to slice the data. For example, a demographic report might indicate that of all the registered users causing events for a given term (i.e., searching using that term), X% are women, Y% are within the ages of 18-25, etc.

4.b. Selling Advertising Space Based on Categorizations and/or Buzz

Categorizing search words has many applications, such as selling advertising space on search page results for searches on a large number of words. This would allow a car manufacturer to specify that their advertisement be shown whenever a search phrase is categorized in a car category. For example, if a visitor searches for "Dodge" and previous user behavior (over possibly many users) had indicated that "Dodge" can be categorized in an automobile category, the advertisement would be shown.

Another use of buzz in relation to advertisements is an application that generates the text and/or other creative components of the advertisement and does so as a function of the top buzz subjects or products for a category of interest to a visitor to a Web site. For example, if a visitor to a site demonstrates interest in "rap music", the application would generate an advertisement that took into account the top buzz for a rap band, such as generating an advertisement that highlighted the offerings of that top rap band.

4.c. Campaign Monitoring

Campaign tracking allows users to measure the impact of their marketing campaigns on generating online buzz. Fig. 13 illustrates a basic campaign monitoring page. Pre-campaign buzz can be compared with buzz during and after the campaign, as shown in Fig. 14.

4.d. Intersection Analysis

Fig. 15 illustrates intersection analysis. Intersection analyses of the demographics of users searching for two terms allows users to identify any overlaps between groups of users searching for multiple terms or brands (e.g., Ford and GM, or Britney Spears and Christina Aguilera).

4.e. Associated Interests Analysis

Fig. 16 illustrates associated interests analysis. Associated interests analysis indicates the other interests of users searching for a particular term. For example, of those people searching for Ford, the other terms/categories they are searching for can be tracked.

5. Variations on the Basic System

The above description is intended as a thorough teaching of how to make and use a statistics monitor and several variations. The above description is not intended to be exhaustive of the possibilities. For example, the above description generally assumes that the interconnecting media between the users and the monitored site is the Internet, but the Internet can be replaced with other media without departing from the scope of the invention, such as a non-TCP/IP network, a Local area network (LAN), and intranet, a virtual private network (VPN), or a wireless-access protocol (WAP) network. While the above systems may have been explained with reference to a particular criteria for counting, such as only one count per unique user per day, other criteria might be used, such as incrementing once every time a user causes an event, or once per user per day.

The above description should not be construed to be limiting to particular computing devices, as the statistics monitor might monitor visits by users with WAP devices, handheld computers, embedded computers, laptops computers and Web-enabled devices, to name a few. In a practical system, the monitor might handle multiple types of devices and might even track statistics by device type or track different device types differently.

The pages being viewed need not be HTML, but might be dynamic server pages, ASP pages, for example. Also, the "buzz" is not the only statistic that can be tracked. For example, some other variable can be tracked. In a particular example, results from the statistics monitor might be used to calculate a charge to the user where the page views are not free but are related in some way to the statistical results.

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.